Robust Incomplete-Modality Alignment for Ophthalmic Disease Grading and Diagnosis via Labeled Optimal Transport

Qinkai Yu[1], Jianyang Xie[2], Yitian Zhao[3], Cheng Chen[4], Lijun Zhang[5], Liming Chen[6], Jun Cheng[7], Lu Liu[1], Yalin Zheng[2], Yanda Meng[1(✉)]

[1] Computer Science Department, University of Exeter, Exeter, UK.
[2] Eye and Vision Sciences Department, University of Liverpool, Liverpool, UK.
[3] Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China.
[4] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China.
[5] Key Laboratory of System Software, Institute of Software, Chinese Academy of Sciences, Beijing, China.
[6] School of Computer Science and Technology, Dalian University of Technology, Dalian, China.
[7] Institute for Infocomm Research, A*STAR, Singapore.
Y.M.Meng@exeter.ac.uk

Abstract. Multimodal ophthalmic imaging-based diagnosis integrates colour fundus imaging with optical coherence tomography (OCT) to provide a comprehensive view of ocular pathologies. However, the uneven global distribution of healthcare resources often results in real-world clinical scenarios encountering incomplete multimodal data, which significantly compromises diagnostic accuracy. Existing commonly used deep learning pipelines, such as modality imputation and distillation methods, face notable limitations: (1) Imputation methods struggle with accurately reconstructing key lesion features, since OCT lesions are localized, while fundus images vary in style. (2) distillation methods rely heavily on fully paired multimodal training data. To address these challenges, we propose a novel multimodal alignment and fusion framework capable of robustly handling missing modalities in the task of ophthalmic diagnostics. By considering the distinctive feature characteristics of OCT and fundus images, we emphasise the alignment of semantic features within the same category and explicitly learn soft matching between modalities, allowing the missing modality to utilize existing modality information, achieving robust cross-modal feature alignment under the missing modality. Specifically, we leverage the Optimal Transport (OT) mechanism for multi-scale modality feature alignment: class-wise alignment through predicted class prototypes and feature-wise alignment via cross-modal shared feature transport. Furthermore, we propose an asymmetric fusion strategy that effectively exploits the distinct characteristics of OCT and fundus modalities. Extensive evaluations on three large-scale ophthalmic multimodal datasets demonstrate our model's superior performance under various modality-incomplete scenarios, achiev-

ing state-of-the-art performance in both complete modality and inter-modality incompleteness conditions. The implementation code is available at https://github.com/Qinkaiyu/RIMA.

Keywords: Ophthalmic Missing Modality · Optimal Transport.

1 Introduction

Retinal fundus imaging and Optical Coherence Tomography (OCT) serve as prevalent two and three dimensional imaging modalities for diagnosing a wide range of ophthalmic conditions. In clinical diagnostics settings, fundus imaging is used to assess the distribution of lesions on the retinal surface, while OCT provides cross-sectional scans for detailed examination of more profound retinal abnormalities [19]. These two imaging modalities complement each other and differ in anatomy appearance and informational depth. For instance, fundus imaging vividly displays vascular structures and various pathological observations on the retinal surface but cannot reflect the difference in retinal layer thickness. Conversely, OCT can objectively evaluate retinal thickness and subretinal lesions but cannot visualize retinal blood vessels or other essential retinal features [2]. Compared to single-modality approaches, multimodal diagnostic strategies provide a more comprehensive and accurate assessment of ophthalmic conditions [19] via integrating fundus and OCTs. However, in real-world clinical settings, challenges such as high equipment costs, time constraints often result in incomplete data from one of the modalities.

Currently, commonly used methods to address the issue of missing modalities can be categorized into two main types: 1) modality imputation [14,22] and 2) distillation methods [12,7]. modality imputation methods fill in missing modality samples or generate the missing modality by performing various transformations on the existing modality, either at the image level or the representation level. However, the modality imputation mechanism for OCT and fundus imaging is limited by the difficulty of accurately reconstructing and aligning key lesion features in the presence of missing modalities. This is because the OCT data may not contain sufficient semantic lesion information due to the natural gaps between slices (*e.g.,* 6 mm gaps between each slice) [4]. Differently, fundus images exhibit diverse stylistic features among samples due to varying imaging conditions, equipment differences, and pathological changes. These factors significantly impair the ability of modality imputation methods to accurately generate critical semantic features, leading to severe overfitting problems. distillation methods employ distillation strategies [7] to ensure that the modality fusion process is not constrained by the missing ones. However, these methods rely heavily on complete multimodal data and struggle to manage incomplete modalities during training. Specifically, these methods customize architectures tailored to specific datasets and assume that a complete dataset is available during training while missing modalities only occurred during testing [20]. In the cases where OCT data is limited, these approaches are unable to effectively utilize the scarce OCT information to improve diagnostic performance. Therefore,

there is a need to develop diagnostic methods that can fully utilize the multimodal information from OCT and fundus imaging while maintaining robust performance in the face of totally missing modality data.

In this work, we propose a novel multi-modal alignment and fusion paradigm based on Optimal Transport (OT) for ophthalmic disease diagnosis (overview shown in Figure 1) that can efficiently and robustly learn multimodal feature representations in various modality incomplete and complete scenarios. Specifically, OT is employed to achieve class-wise and feature-wise alignment under missing modality conditions, building a bridge between the two modalities so that the missing modality can utilize the information from the existing modality. OT serves as a computational framework for cross-modal alignment, achieving optimal matching with minimal overall cost [3,1,17,11]. The contributions of this work are summarized as follows: (1) We propose class-wise and feature-wise alignment based on Labeled OT. Different from existing studies [16,21], our method neither relies on strict one-to-one alignment nor performs global alignment between the two modalities. Instead, by considering the unique features of OCT and fundus images, we focus the alignment on semantic features within the same class and explicitly learn soft matching between modalities, thereby achieving more robust cross-modal alignment. (2) Additionally, we propose an asymmetric fusion strategy tailored to address the feature discrepancies between OCT and fundus images. (3) We comprehensively evaluate the proposed method on three large-scale ophthalmic multi-modal datasets under various modality-incomplete and complete scenarios. The results demonstrate that our method achieves state-of-the-art performance.

2 Methods

2.1 Labeled OT Match

Let $\{x, y\} = \{x_n, y_n\}_n^N$ as a multimodal dataset that has N samples. Each $x_n = \{x_n^f, x_n^o\}$ consists of 2 inputs from fundus and OCT modalities, respectively. And $y_n \in 1, 2, ..., C$ where C is the number of categories. The modality encoder learn from the input modality x^f and x^o to produce the single-modal representation $e^f \in \mathbb{R}^{N \times D_f}$ with distribution μ and $e^o \in \mathbb{R}^{N \times D_o}$ with distribution ν. $\mathcal{T}_{\mu,\nu} = \{T \in \mathbb{R}_+^{N,N} | T\mathbb{1}_N = \mu, T^\top \mathbb{1}_N = \nu\}$ represent the set of feasible transport plans. $T \in \mathbb{R}_+^{N,N}$ involving the marginal constraints of total mass equality between two distributions μ and ν. The Gromov-Wasserstein Optimal Transport (GWOT) [9], which leverages the relationships between intra-domain distances as its cost function, is deemed more appropriate for cross-modal alignment [10,11]. Differently, we introduce constraints to limit transport operations within each class, aligning identical lesions across modalities. Specifically, we employed Labeled OT based on the Gromov-Wasserstein distance, ensuring that the optimization process remains consistent with GWOT. The additional constraints refine the $\mathcal{T}_{\mu,\nu}$ by incorporating both marginal constraints and class-specific restrictions, as defined below:

$$\bar{\mathcal{T}}_{\mu,\nu} = \{T \in \mathbb{R}_+^{N,N} | T \in \mathcal{T}_{\mu,\nu}, T_{ij} > 0 \Rightarrow y_i^f = y_j^o\}. \quad (1)$$

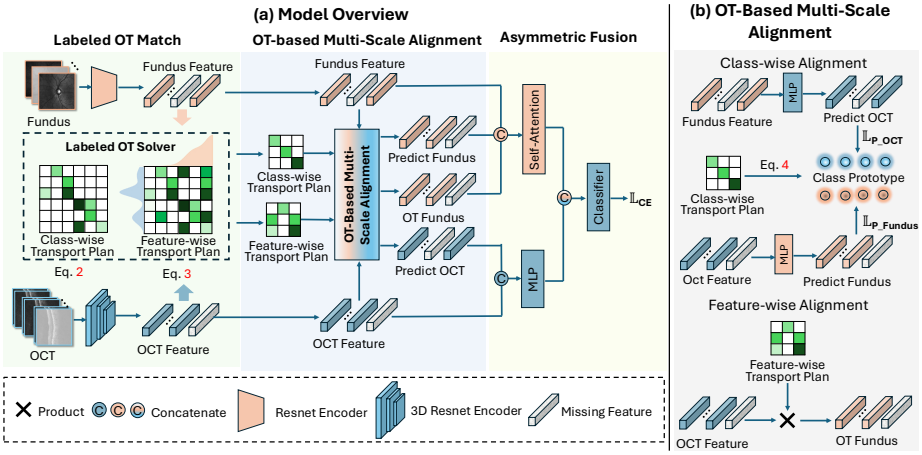

Fig. 1: Overview of the proposed framework. (a) The framework consists of three parts: Labeled OT Match, OT-based Multi-Scale Alignment, and Asymmetric Fusion. We leverage the matching results from Optimal Transport to construct class prototypes and cross-modal shared features for multi-scale alignment. To accommodate the distinct characteristics of each modality, we introduce an asymmetric fusion strategy to adaptively integrate modality-specific information. (b) Details of the Multi-Scale Alignment module, showing both class-wise and feature-wise alignment.

Using from OCT to fundus alignment as an example, Labeled GWOT process aim to find an optimal transport plan $T^* \in \bar{\mathcal{T}}_{\mu,\nu}$ by sinkhorn iterations [13], the optimization process for the class-wise transport plan is defined as:

$$\min_{T_c \in \bar{\mathcal{T}}_{\mu,\nu}} \sum_{i,j,k,l} \underbrace{|(e_i^f - e_k^f)^2 - (e_j^o - e_l^o)^2|^2}_{GW\,Distance} \times T_{c_{i,j}} T_{c_{k,l}} - \epsilon H(T_c), \tag{2}$$

where $\epsilon H(T_c)$ is the entropy regularization with the scaling factor ϵ. Here, we define the optimal transport plan optimize by Equation 2: (1) from fundus to OCT class-wise transport as T_c^{fo} and (2) from OCT to fundus class-wise transport plan as T_c^{of}. Then T_c^{of} is used for OCT to fundus feature-wise transport and the optimization process for the feature-wise transport plan $T_v \in \mathbb{R}_+^{D_o \times D_f}$ is defined as:

$$\min_{T_v} \sum_{i,j,k,l} \underbrace{|(e_i^f - e_k^f)^2 - (e_j^o - e_l^o)^2|^2}_{GW\,Distance} \times T^{of}_{c_{i,j}} T_{v_{k,l}} - \epsilon H(T_v). \tag{3}$$

Notably, as for the OCT-to-fundus alignment, we apply both class-wise and feature-wise transport. Meanwhile, for the fundus-to-OCT alignment, only class-wise transport is used. This distinction arises because fundus imaging offers a

more comprehensive global perspective compared to the OCT data, making feature transport alignment more effective in capturing global structural information and semantic features within the fundus modality. Conversely, the OCT modality is subject to higher levels of noise, and the lesions are concentrated in localized regions. Therefore, relying primarily on class transport ensures the semantic alignment of corresponding lesions while avoiding additional noise.

2.2 OT-based Multi-Scale Alignment

Class-wise Alignment. Here, we illustrate the fundus-to-OCT alignment as an example. Element $T^{fo}_{c_{i,j}}$ denotes the probability of transferring e^f_i to e^o_j. And $T^{fo}_{c_{i,\cdot}}$ indicates the distribution over potential matches for the fundus representation e^f_i. We sample matching ground truth e^o_j for e^f_i according to the probability distribution of potential matches: $p(j|i) = \frac{T^{fo}_{c_{i,\cdot}}}{\sum_{j'} T^{fo}_{ci,j'}}$. As the training progresses over multiple epochs, the collection of all e^o_j that is actually matched with e^f_i, can be interpreted in expectation as a 'soft prototype' (c^o_i), which is defined as:

$$c^o_i = \mathbb{E}[e^o_j | e^f_i] = \sum p(j|i) e^o_j. \tag{4}$$

In other words, each e^f_i (or a batch of similar e^f_i) is associated with a weighted combination of possible matches that are determined through repeated random sampling. This stands in contrast to traditional clustering methods that typically employ a single (or a few) shared prototypes and strict one-to-one mappings. This design captures intra-class diversity while maintaining global consistency in the cross-modality alignment. We further train MLPs and treat each sample's class prototype as the ground truth to achieve cross-modal class prototype alignment. We use cosine similarity as the objective function for fundus to OCT ($\mathbb{L}_{\text{p\_OCT}}$) and OCT to fundus ($\mathbb{L}_{\text{p\_fundus}}$):

$$\mathbb{L}_{\text{p\_OCT}} = 1 - \sum_{i=1}^{n} \frac{e^f_i c^o_i}{\|e^f_i\| \|c^o_i\|}, \qquad \mathbb{L}_{\text{p\_fundus}} = 1 - \sum_{i=1}^{n} \frac{e^o_i c^f_i}{\|e^o_i\| \|c^f_i\|}, \tag{5}$$

where primarily to promote semantic alignment in high-dimensional space rather than merely matching vector magnitudes.

Feature-wise Alignment. The feature level alignment is only used for OCT to fundus modality alignment. This is mainly due to the characteristics of OCT modality, that the lesion areas in OCT are concentrated and localized, while the fundus modality provides a global perspective, making such feature-wise global alignment only applicable to the fundus modality. During the alignment process from fundus to OCT, feature alignment is prone to overfitting to noisy patterns in OCT data, resulting in a decrease in the model's generalization performance. Given transport plan T_v, the OT fundus is identified by: OT fundus $= T_v^\top e^o$. OT fundus aligns the fundus feature distribution with OCT feature distribution, thereby preserving the latent structure across modalities and providing a global perspective across modalities in favour of the fundus modality.

Table 1: Comparison with other methods. The performance of other models is from *Liu et al.* [7], where our model follows the same experimental settings. The best results are **bolded**, with underline indicating the second highest.

Method	Modality		Harvard-30k AMD			Harvard-30k DR			Harvard-30k Glaucoma			Average		
	OCT	fundus	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Inter-Modality Missing with ResNet-50 Backbone														
2D-Resnet-50		✓	73.12	75.35	72.23	73.81	79.15	70.47	73.12	75.35	72.23	73.35	76.61	71.64
B-CNN + distill		✓	72.35	71.98	70.03	73.62	67.50	69.68	73.39	76.61	72.47	73.12	72.03	70.72
M^{2}LC + distill		✓	73.24	72.67	73.80	73.04	67.89	74.59	72.78	70.23	71.11	73.02	70.26	73.16
IMDR		✓	75.17	80.48	**76.59**	76.19	79.07	72.18	75.54	78.47	**75.12**	75.63	79.34	74.63
Ours		✓	**75.83**	**83.75**	74.46	**82.74**	**85.39**	**82.70**	**75.72**	**78.81**	74.76	**78.09**	**82.65**	**77.30**
3D-Resnet-50	✓		65.17	69.98	69.63	70.73	69.94	61.97	65.72	69.89	70.98	67.20	69.93	67.52
B-CNN + distill	✓		69.57	70.14	67.45	69.05	65.25	67.93	69.64	68.95	67.18	69.42	68.11	67.52
M^{2}LC + distill	✓		68.97	72.23	65.06	67.20	65.05	64.33	67.70	71.22	65.60	67.95	69.50	64.99
IMDR	✓		70.62	72.69	**71.90**	**72.62**	**74.69**	**72.90**	71.16	75.07	70.37	71.46	74.15	71.72
Ours	✓		**74.50**	**79.21**	70.59	72.02	73.56	71.81	**75.47**	**80.94**	**75.13**	**73.99**	**77.90**	**72.51**
Complete-Modality Fusion with ResNet-50 Backbone														
B-CNN	✓	✓	71.67	81.22	69.01	73.55	75.44	73.96	73.66	79.37	72.89	72.96	78.67	71.95
B-EF	✓	✓	69.33	79.12	60.64	74.98	74.13	77.17	74.19	77.52	72.22	72.83	76.92	70.01
CR-AF	✓	✓	73.00	81.37	67.73	76.57	78.35	76.27	73.92	77.48	73.09	74.49	79.06	72.36
M^{2}LC	✓	✓	74.00	**84.56**	70.73	77.36	83.62	76.01	74.73	77.45	73.54	75.36	81.87	73.30
Eye-Most	✓	✓	75.55	82.30	71.00	75.00	80.34	73.42	74.33	75.40	72.01	74.96	79.34	72.08
IMDR	✓	✓	**76.45**	83.10	73.57	78.03	85.34	78.23	**76.50**	77.11	**74.36**	76.99	81.85	75.38
Ours	✓	✓	75.17	83.20	**74.33**	**82.14**	**86.48**	**80.70**	74.39	**81.36**	73.72	**77.23**	**83.68**	**76.25**

2.3 Asymmetric Fusion

We designed asymmetric fusion to address the differences in characteristics between fundus and OCT modalities. For the fundus modality, we concatenate three types of features(the cross-modality feature from feature-wise alignment, the semantic consistency features produced by class-wise alignment and the feature from backbone) and feed them into an attention layer. The three types of features represent the correspondences between vascular structure and OCT stratification, lesion category priors, and local details representation, respectively. The attention layer adaptively learns nonlinear correlations between features through dynamic weight allocation, avoiding the inductive bias introduced by manually designed fusion rules [15]. For OCT modality, semantic consistency features after class-wise alignment and the feature extracted by the backbone are fed into MLPs. This design is based on the susceptibility of OCT images to motion artifacts and scattering noise interference, avoiding the overfitting of noise patterns in the training dataset by complex fusion modules. The two fused features are concatenated and fed into the classification head. The total loss function (L_{total}) integrates the classification loss and the alignment losses together:

$$L_{total} = \mathbb{L}_{\text{ce}} + \mathbb{L}_{\text{p\_OCT}} + \mathbb{L}_{\text{p\_fundus}}, \tag{6}$$

where $\mathbb{L}_{ce}$ is a cross-entropy loss for the ophthalmic disease diagnosis task. The hyper-parameter each term's weight is set as 1 by default.

3 Experiments

Datasets and Evaluation. We evaluate the proposed framework using three dataset: Harvard-30k AMD, Harvard-30k DR, and Harvard-30k Glaucoma from the publicly available multimodal datasets Harvard-30k [8], covering Age-related

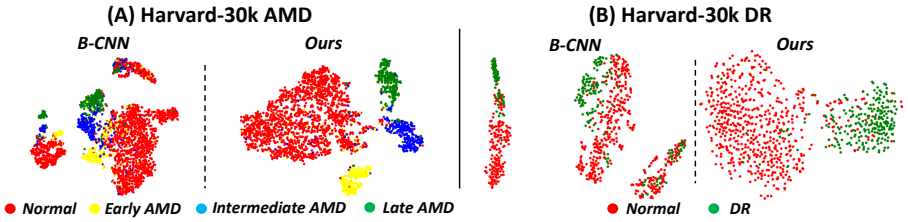

Fig. 2: A t-SNE visualization of the feature distribution from B-CNN and our model, captured before the final FC layer on the AMD and DR datasets.

Macular Degeneration (AMD), Diabetic Retinopathy (DR), and Glaucoma diseases. Harvard-30k subsets are annotated with four-tier AMD grading and two-tier for glaucoma and DR, with dimensions of 448 × 448 for fundus and 200 × 256 × 256 for OCT. To ensure reliable results, each dataset underwent five-fold cross-validation. Performance was evaluated using Accuracy (ACC), F1-Score (F1), Area Under the Curve (AUC).

Implementation Details. All compared models use ResNet50 [5] as the backbone for both OCT and fundus modality. The input fundus images were resized to 384 × 384; OCT images were resized to 96 × 96 × 96. All of the models were trained with a batch size of 32 for the same number of epochs.

Experiments Setting. To comprehensively evaluate the performance of the proposed models, we designed the following three experimental setups: 1) **Complete-Modality Fusion:** Train on the complete training set and evaluate on the complete testing set; 2) **Inter-Modality Missing:** Train on a complete training set and evaluate on test set with single modality (OCT or fundus); 3) **Proportional Random Missing:** Train on the training set with a partial proportion of missing modalities and evaluate on the testing set with a partial proportion of missing modalities (same missing ratios for training and test sets).

Complete-Modality Fusion. we compare our model with six representative complete-modality fusion methods, including B-CNN(baseline model utilizing an intermediate multimodal fusion method), B-EF [6], CR-AF [23], M^{2}LC [18], Eye-Most [24], IMDR [7]. As shown in Table 1, our model consistently outperforms other methods across all datasets and backbones evaluated. For instance, our model outperforms Eye-most [24] 5.8% on F1 and 5.4% on AUC on average and exceeds IMDR [7] 1.1% on F1 and 2.2% AUC on average. To illustrate the effects of our model, we perform a t-SNE visualization (shown in Figure 2) on AMD and DR testset under Complete-Modality Fusion.

Inter-Modality Missing. As shown in Table 1, Compared to the evaluation strategy with complete modality fusion, the performance of all other models declines across different datasets when modality is missing. In contrast, our model maintains outstanding performance, exceeding IMDR [7] 4% AUC onaverage when OCT is missing and 5% when fundus is missing. Notably, while the OCT data in the AMD and Glaucoma datasets manifests highly distinct lesion characteristics that are readily discernible by medical professionals, those in the DR

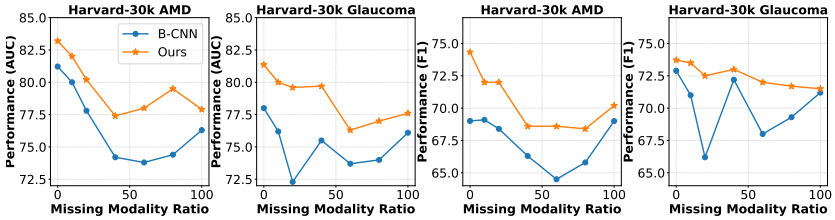

Fig. 3: Comparison of B-CNN and Ours in proportional random missing experiment, evaluated across various missing rates on different ophthalmic diseases and datasets.

Table 2: Ablation study of each components on Harvard-30k DR dataset.

Variants	OCT	fundus	ACC	AUC	F1
w/o Class-wise Alignment		✓	77.38	76.20	81.42
Ours		✓	**82.74**	**85.39**	**82.70**
w/o Class-wise Alignment	✓		70.24	68.14	70.88
w/o Feature-wise Alignment	✓		70.83	73.11	70.69
Ours	✓		**72.02**	**73.56**	**71.81**
w/o Asymmetric Fusion	✓	✓	79.17	85.42	77.80
Ours	✓	✓	**82.14**	**86.48**	**80.70**

dataset are relatively inconspicuous. The diagnostic performance of our model in the event of fundus modality absence is highly congruent with the aforementioned phenomena. This unique trait, however, remains unobserved in other models. This compellingly attests to the fact that our model can precisely and effectively align the semantic features of lesions.

Proportional Random Missing. As shown in Figure 3, by randomly masking a portion of the modality data from 0% to 100%, we compared our model with B-CNN in the AMD and Glaucoma datasets. Our model performs outstandingly and demonstrates extremely strong robustness, maintaining relatively stable performance compared to the baseline models. This indicates that our proposed model effectively aligns the key lesion features and reconstructs information under missing modalities setting, equipping it to handle interferences in real-world scenarios.

Ablation Study of Proposed Components. To assess the effectiveness of our model, we conducted a thorough ablation study on the DR dataset, shown in Table 2. We evaluated the performance of the model by removing the Class-wise Alignment, Feature-wise Alignment, and Asymmetric Fusion modules individually and remaining the rest of the model structure unchanged. Notably, removing the Class-wise Alignment has the most significant impact on AUC (7% for fundus modality, 8% for OCT modality). Feature-wise Alignment and Asymmetric Fusion modules also contribute to the improvement of the DR grading effect, such as 0.6% and 1.2% of AUC. These experimental results demonstrate the effectiveness of each component based on different modality properties.

4 Conclusion

We proposed a novel multimodal alignment and fusion framework. Our model, based on Labeled OT, has achieved multi-scale alignment at both the class and feature levels. This ensures precise matching of key lesion features to obtain robust and effective multimodal feature representations while maintaining excellent performance even in the presence of missing modalities. Furthermore, we fully consider the characteristics of different modalities and design an asymmetric fusion mechanism to maximize the advantages of each modality. Experimental results on three datasets demonstrate that our method significantly outperforms state-of-the-art models under different experimental settings. In particular, our approach has exhibited exceptional robustness to missing modalities, further validating its effectiveness across different ratio of modality missing in real-world scenarios.

Acknowledgments This work is supported by Y. Meng's The Royal Society Fund (IEC\NSFC\242172), United Kingdom. This work is also partly funded by the UK Engineering and Physical Sciences Research Council [grant number EP/X01441X/1] and Medical Research Council (grant number MR/Z506175/1). Qinkai Yu thanks the PhD studentship funded by Liverpool Centre for Cardiovascular Science and University of Exeter.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, Z., Xu, Q., Yang, Z., He, Y., Cao, X., Huang, Q.: Otkge: Multi-modal knowledge graph embeddings via optimal transport. Advances in Neural Information Processing Systems **35**, 39090–39102 (2022)
2. De Carlo, T.E., Romano, A., Waheed, N.K., Duker, J.S.: A review of optical coherence tomography angiography (octa). International journal of retina and vitreous **1**, 1–15 (2015)
3. Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., Chilimbi, T.: Multi-modal alignment using representation codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15651–15660 (2022)
4. Fujimoto, J., Swanson, E.: The development, commercialization, and impact of optical coherence tomography. Investigative ophthalmology & visual science **57**(9), OCT1–OCT13 (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hua, C.H., Kim, K., Huynh-The, T., You, J.I., Yu, S.Y., Le-Tien, T., Bae, S.H., Lee, S.: Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images. IEEE Journal of Biomedical and Health Informatics **25**(7), 2686–2697 (2020)
7. Liu, c., Huang, z., Chen, Z., Tang, F., Tian, Y., Xu, Z., Luo, Z., Zheng, Y., Meng, Y.: Incomplete modality disentangled representation for ophthalmic disease grading and diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6225–6233 (2025)

8. Luo, Y., Tian, Y., Shi, M., Elze, T., Wang, M.: Eye fairness: A large-scale 3d imaging dataset for equitable eye diseases screening and fair identity scaling. arXiv preprint arXiv:2310.02492 (2023)
9. Peyré, G., Cuturi, M., Solomon, J.: Gromov-wasserstein averaging of kernel and distance matrices. In: International conference on machine learning. pp. 2664–2672. PMLR (2016)
10. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
11. Ryu, J., Bunne, C., Pinello, L., Regev, A., Lopez, R.: Cross-modality matching and prediction of perturbation responses with labeled gromov-wasserstein optimal transport. arXiv preprint arXiv:2405.00838 (2024)
12. Saha, P., Mishra, D., Wagner, F., Kamnitsas, K., Noble, J.A.: Examining modality incongruity in multimodal federated learning for medical vision and language-based disease detection. arXiv preprint arXiv:2402.05294 (2024)
13. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics **21**(2), 343–348 (1967)
14. Sun, Y., Liu, Z., Sheng, Q.Z., Chu, D., Yu, J., Sun, H.: Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion **110**, 102454 (2024)
15. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
16. Wang, D., Li, M., Liu, X., Xu, M., Chen, B., Zhang, H.: Tuning multi-mode token-level prompt alignment across modalities. Advances in Neural Information Processing Systems **36** (2024)
17. Wang, Y., Zhang, T., Zhang, X., Cui, Z., Huang, Y., Shen, P., Li, S., Yang, J.: Wasserstein coupled graph learning for cross-modal retrieval. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1793–1802. IEEE (2021)
18. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
19. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al.: Gamma challenge: glaucoma grading from multi-modality images. Medical Image Analysis **90**, 102938 (2023)
20. Wu, R., Wang, H., Chen, H.T., Carneiro, G.: Deep multimodal learning with missing modality: A survey. arXiv preprint arXiv:2409.07825 (2024)
21. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21241–21251 (2023)
22. Yang, Y., Chen, H., Chang, Z., Xiang, Y., Ye, C., Ma, T.: Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In: Medical Imaging With Deep Learning. pp. 1006–1018. PMLR (2024)
23. Zheng, J., Liu, H., Feng, Y., Xu, J., Zhao, L.: Casf-net: Cross-attention and cross-scale fusion network for medical image segmentation. Computer Methods and Programs in Biomedicine **229**, 107307 (2023)
24. Zou, K., Lin, T., Han, Z., Wang, M., Yuan, X., Chen, H., Zhang, C., Shen, X., Fu, H.: Confidence-aware multi-modality learning for eye disease screening. Medical Image Analysis **96**, 103214 (2024)